Point Pattern and Univariate point-referenced data analysis of motor vehicle collisions in New York City

Introduction

Since July 2012 the NYPD is releasing collision data on the NYC Open Data portal. It consists of the number of persons being injured and killed, all factors contributing to the collision, time, date, zip-code, and location of the each accident. During this period, almost 90000 people were injured in motor vehicle crashes in New York City. The data from that time through November 2016 was taken and part of it was analyzed. The objectives of this section are to (1) identify if the collisions caused by some factor are clustered; (2) fit the spatial model to see if the location helps to predict the number of persons injured.

Table 1 gives the several rows of the data set. We choose five environmental contributing factors to analyze, Table 2 shows the number of persons injured in the collisions caused by each of them.

Point Pattern Processes

Figure 1 shows the locations of the crashes due to animal action, defective pavement, debris on the road, improper lane marking, and non-working traffic device. According to Figure 2 Manhattan has a much higher density of the crashes than other boroughs, a reflection of its high traffic volume and high daytime population. In addition, Figure 3 indicates that lots of collisions occured in Bronx caused by animals and in Queens due to debris and defective pavement.

We can perform Complete Spatial Randomness test to see if the point patterns are clustered. A point process which is CSR point process is formally defined as a homogeneous

Table 1: NYPD dataset

Latitude	Longitude	# of persons injured	Contributing Factor
40.742 26	−73.977 67	1	Non-Working Traffic Control device
40.857 13	−73.880 79	0	Obstruction/Debris
40.865 21	−73.843 36	2	Animal Action
40.605 45	−73.898 55	2	Pavement Defective
40.817 39	−73.922 76	0	Lane Marking Improper/Inadequate
40.641 96	−73.957 07	0	Pavement Defective
40.708 62	−73.793 95	0	Obstruction/Debris

Table 2: The number of persons injured

Number of persons / Contributing factor	0	1	2	3	4	5	6	8	14	Total
Animal Action	314	36	10	0	0	0	2	1	0	363
Lane Marking Improper	286	26	5	1	0	0	1	0	0	319
Non-Working Traffic Control Device	212	62	23	16	3	0	0	0	0	316
Obstruction/Debris	1134	135	23	8	0	1	0	0	0	1301
Pavement Defective	598	146	32	3	4	0	1	0	1	785
Total	2544	405	93	28	7	1	4	1	1	3084

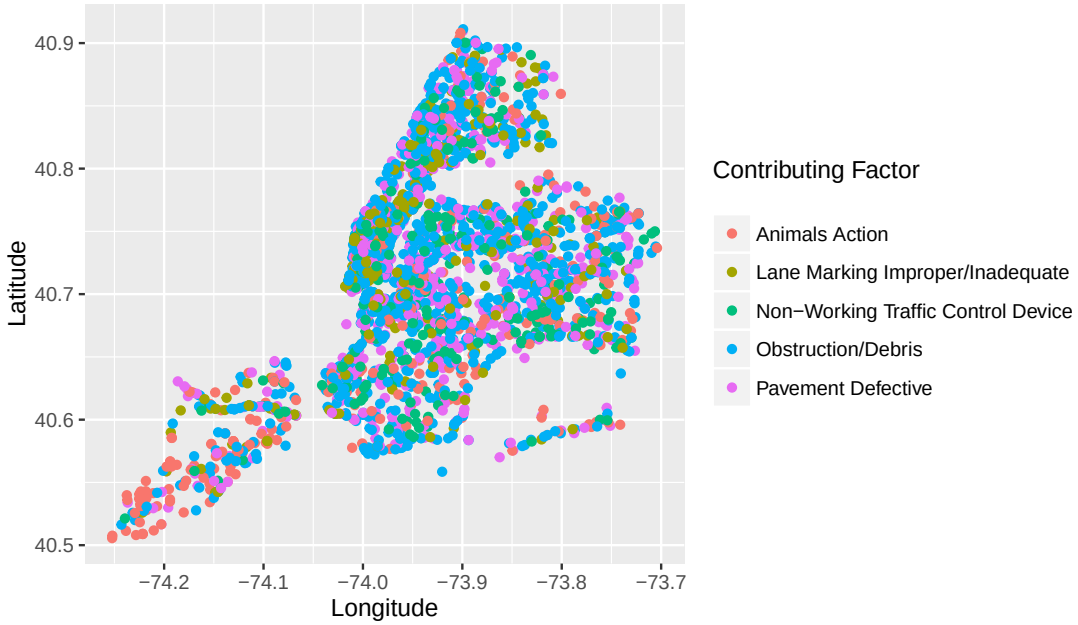

Figure 1: Motor Vehicle Collisions in NYC from July 2012 to November 2016

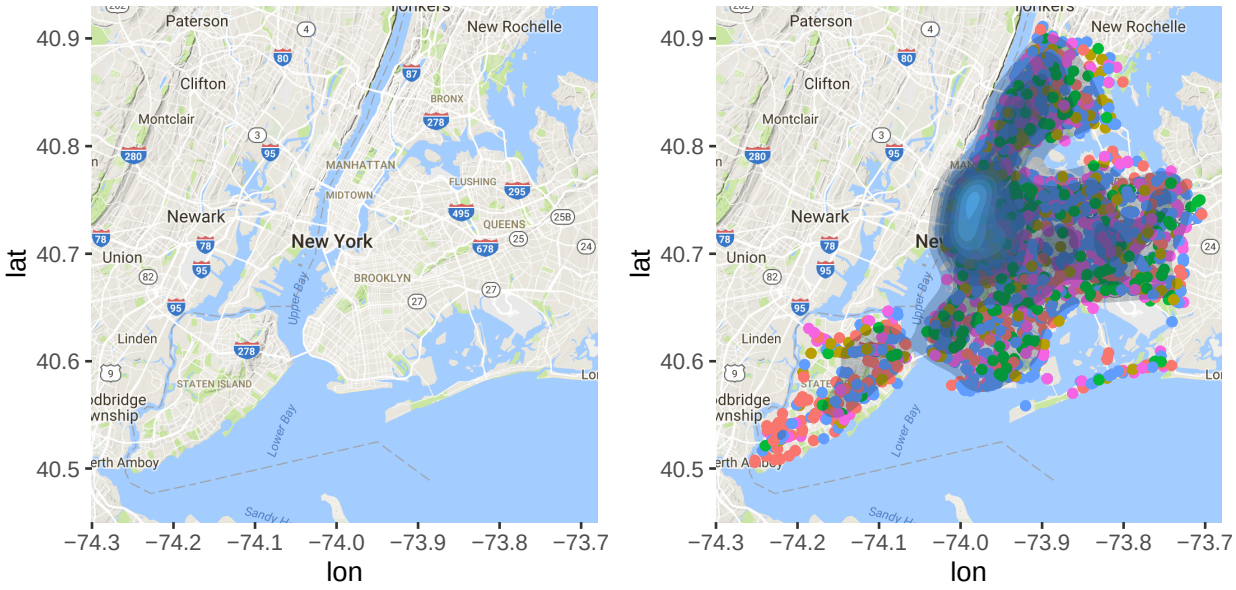

Figure 2: Kernel Density Estimation of Motor Vehicle Collisions in NYC caused by Environmental Contributing Factors

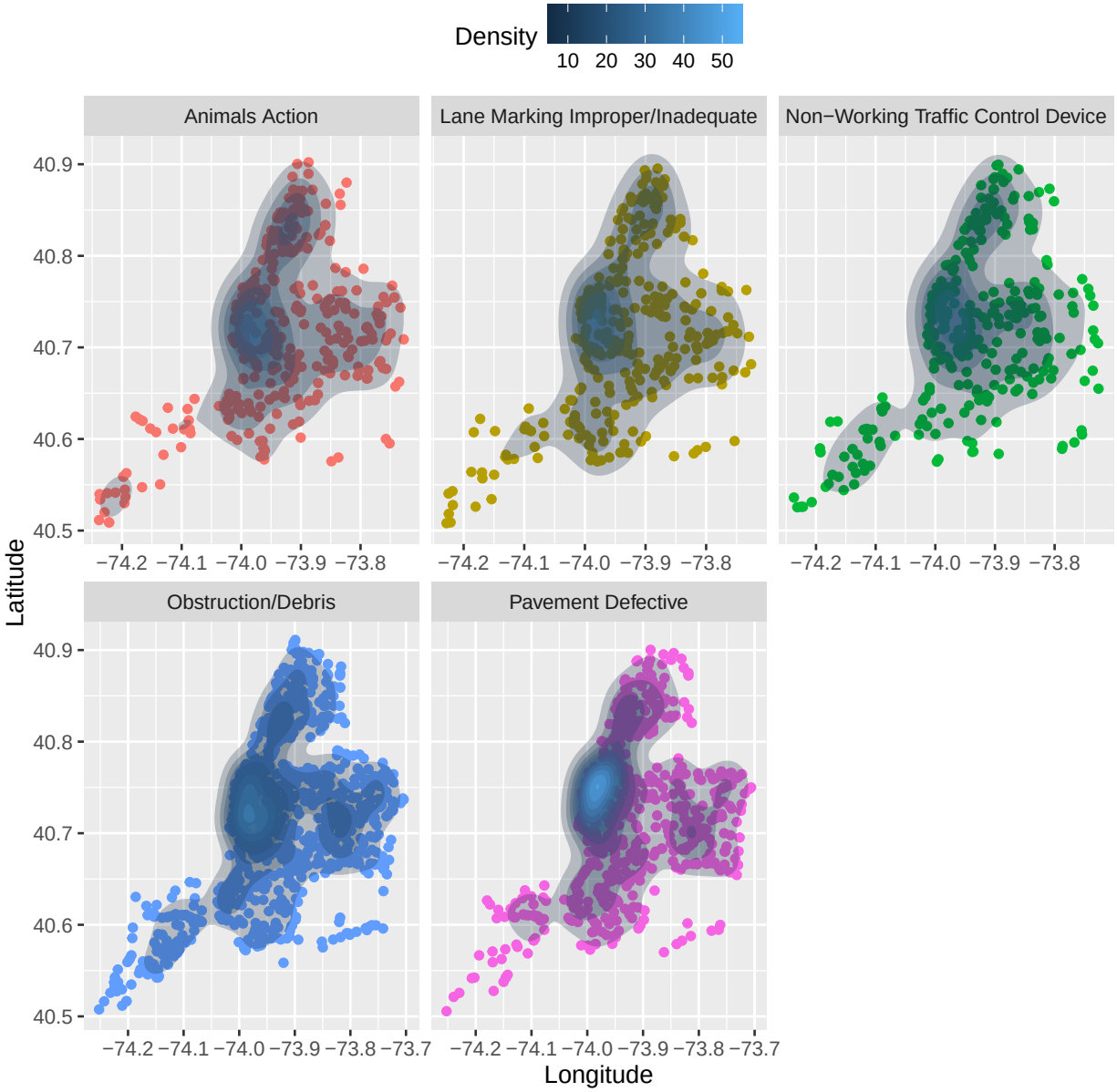

Figure 3: Motor Vehicle Collisions in NYC caused by Environmental Contributing Factors

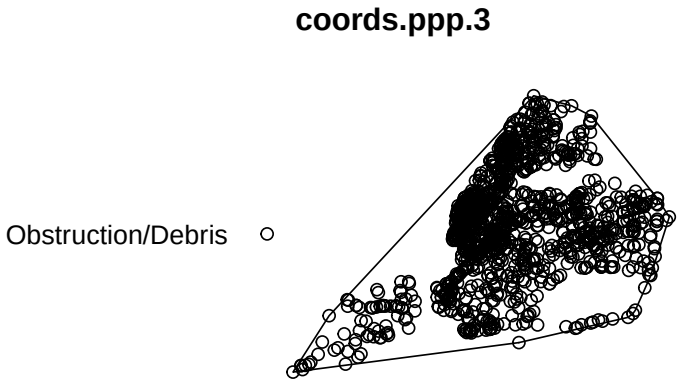

Figure 4: Convex hull of coordinates corresponding to collisions caused by obstruction or debris on the road

Poisson process. Figures 5 and 7 contain theoretical ($G_{pois}, K_{pois}, F_{pois}$) and observed values ($\widehat{G}, \widehat{K}, \widehat{F}$) of the G, K and F functions for point patterns corresponding to the collisions caused by debris and defective pavement respectively. The G function measures the distribution of distances from an arbitrary event to its nearest neighbors. The F function measures the distribution of all distances from an arbitrary point k in the plane to the nearest observed event j. Ripleys K function is the mean number of points per unit area. To calculate these functions we need to create point pattern data (an object of class 'ppp'), for this purpose convex hulls of coordinates were created and shown in Figures 4 and 6. Based on Figures 5 and 7, the lines corresponding to the observed values of G and K functions are above theoretical ones (blue), and the line corresponding to the observed value of the F function is below theoretical value (blue), which indicate that point patterns are clustered. The other three factors produce the same result, showing that the collisions caused by each factor are clustered.

Univariate point-referenced data analysis

Since we are working with big dataset, only two contributing factors (i.e. 'Obstruction/Debris', 'Pavement Defective') were considered, this allows to minimize the computational time of the spLM function. The dataset was splitted into the training (80%) and testing (20%) sets. The ANOVA model was fitted to the training dataset. Based on the Table 3, the contributing factor is significant. Table 4 shows significant positive difference in means between number of persons injured in collisions caused by defective pavement and those by debris on the road. Using testing dataset we found that 95.7% predictive intervals were identified as correct.

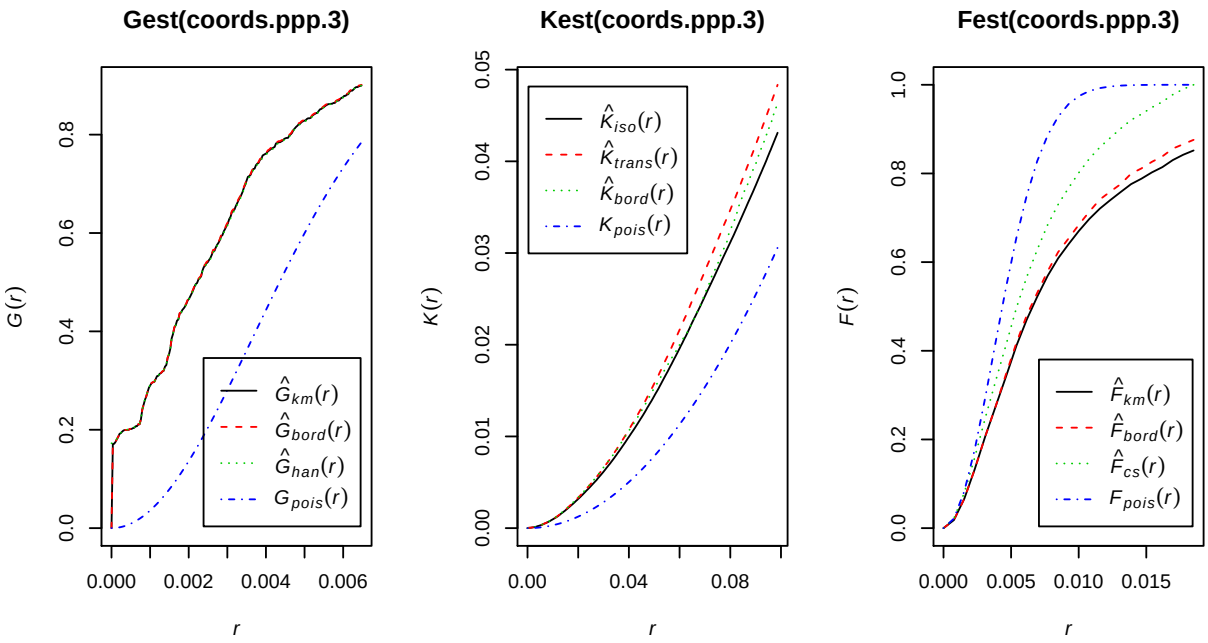

Figure 5: Theoretical and observed values of the G, K and F functions for point pattern corresponding to collisions caused by obstruction or debris on the road

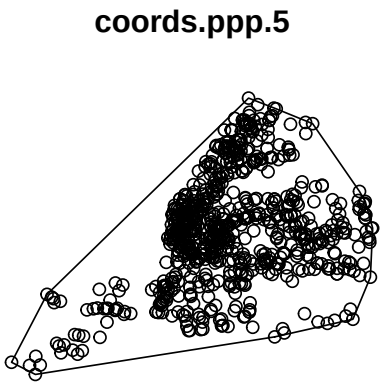

Figure 6: Convex hull of coordinates corresponding to collisions caused by defective pavement

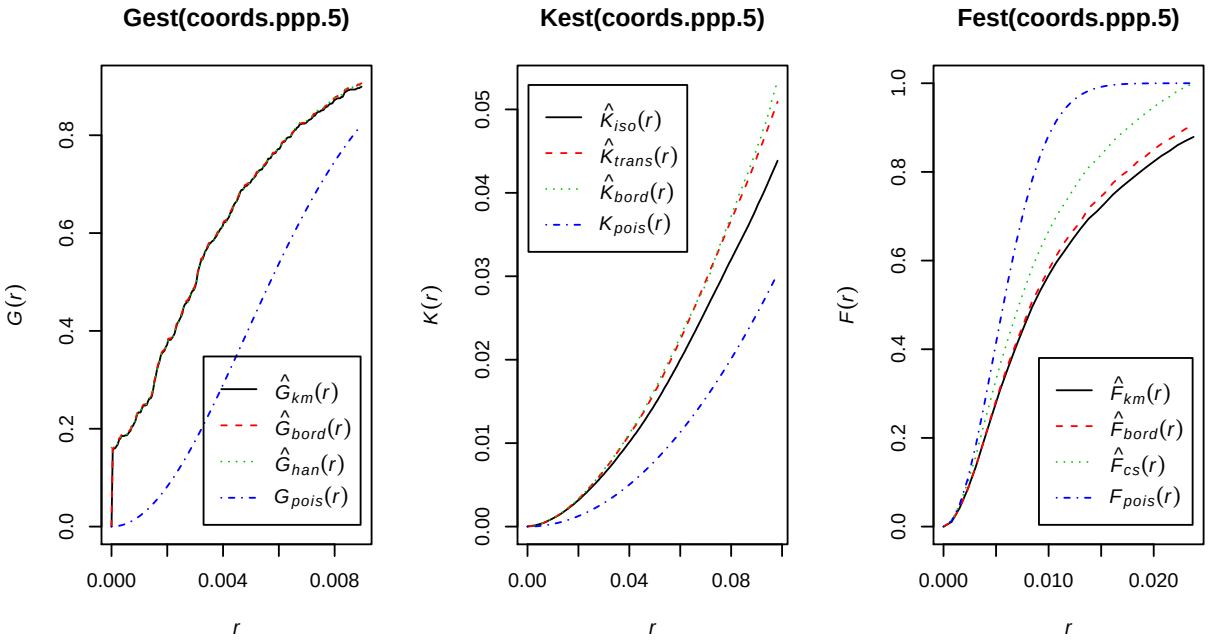

Figure 7: Theoretical and observed values of the G, K and F functions for point pattern corresponding to collisions caused by defective pavement

Source	df	SS	MS	F	Pr(>F)
Contributing factor	1	12.1	12.069	29.48	6.59e-08
Residuals	1487	608.8	0.409		

Table 3: An analysis of variance table

Contributing factor	Diff	Lwr	Upr	P adj
Pavement Defective - Obstruction/Debris	0.1853447	0.1183832	0.2523062	1e-07

Table 4: Tukey pairwise comparison of means

Since our response variable is the number of persons injured we may try to fit poisson generalized linear model, but after checking assumptions, we found that the response does not follow poisson distribution thus we decide to fit univariate gaussian spatial regression model instead.

Figure 8 contains exponential variogram of the residuals of the anova fit. It seems that exponential covariance model fits data well. Based on the graph the estimated nugget, sill, and range are 0.3, 0.6, and 0.35 respectively. We obtained estimates of the partial sill , σ^2, nugget, τ^2, and decay parameter ϕ from the emperical variogram and used them as starting values in the spBayes univariate gaussian spatial regression function spLM. The Uniform(1/60000, 10) prior was used for ϕ, and IG(0.1, 0.1) priors for σ^2 and τ^2. A summary

of the results from the fitted spatial regression model with an exponential covariance function are presented in Table 5, and we see that only the contributing factor is significant. Figure 9 shows trace, density and autocorrelation plots of the posterior samples indicating that chain corresponding to the intercept does not reach a stationarity, but the other chain is mixing well. Figure 10 contains diagnostic plots for two chains with different starting values. The 'potential scale reduction factor' was calculated, the upper confidence limit is 1.09 for the intercept, and 1 for the contributing factor. Since the values which are close to 1 indicating approximate convergence, we can conclude that contributing factor reached stationarity. Using testing dataset we found that 100% predictive intervals were identified as correct, which means that spatial model is working better than anova for the NYPD dataset.

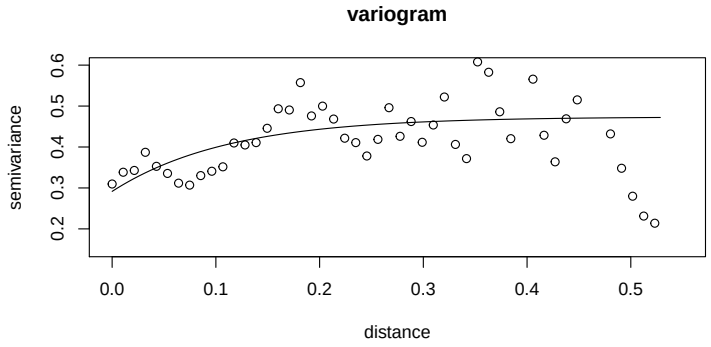

Figure 8: Exponential variogram of the residuals of the anova fit

Table 5: The posterior means, standard deviations, and 2.5% and 97.5% quantiles for each parameter.

Parameter	Mean	SD	2.5% quantile	97.5% quantile
Intercept	0.1703	0. 44793	-0.7449	1.058
Contr. Factor Pavement Defective	0.1849	0.03493	0.1174	0.253

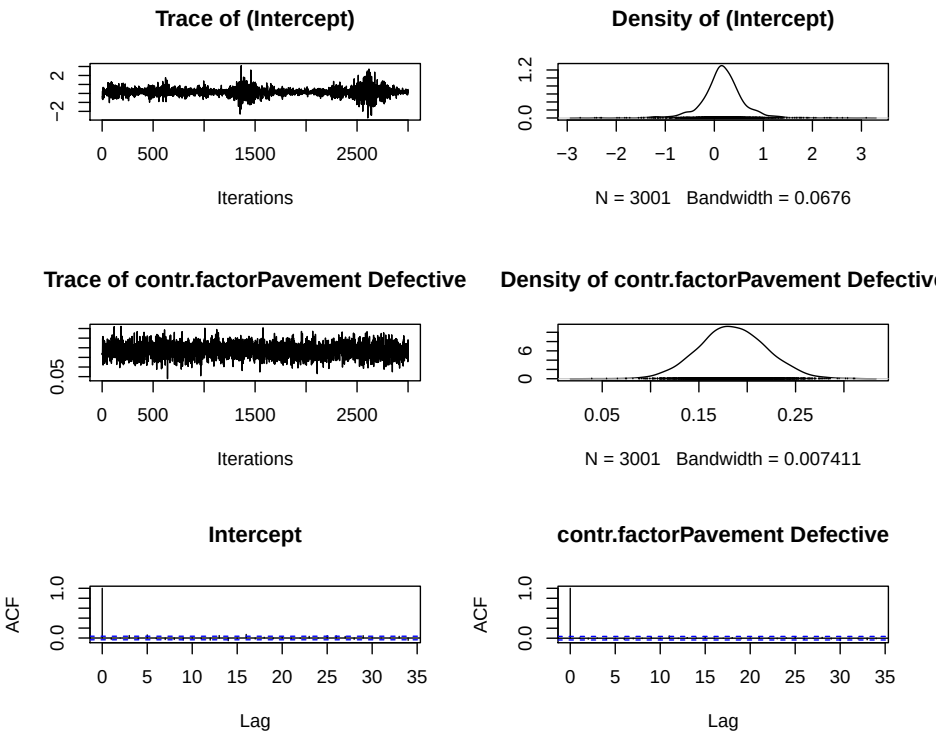

Figure 9: Diagnostic plots from univariate spatial regression model

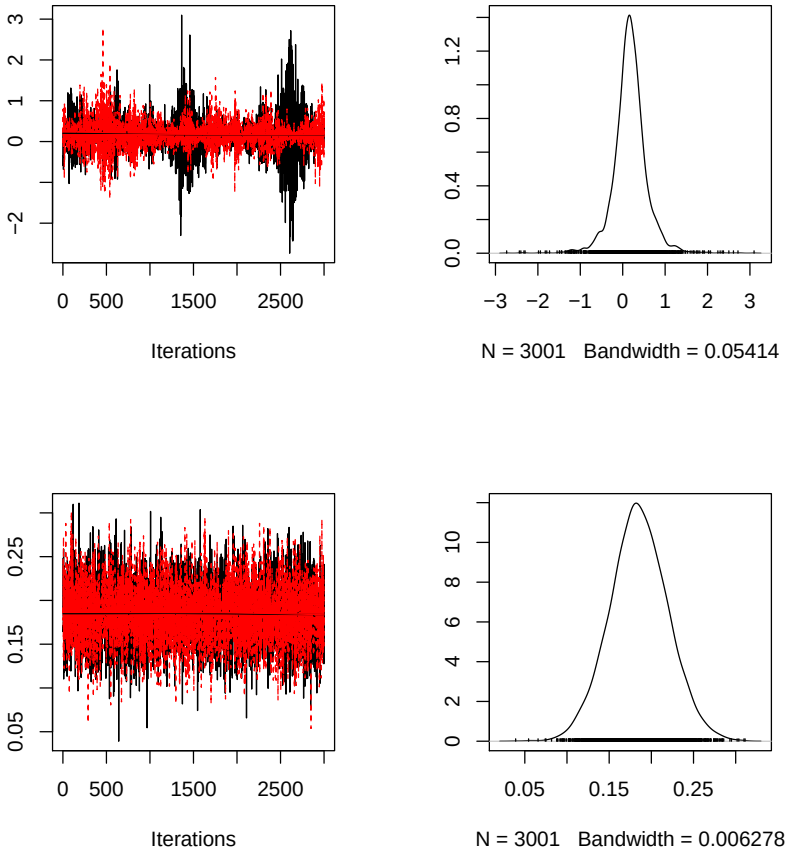

Figure 10: Diagnostic plots for two chains with different starting values